A DataFlux White Paper

Prepared by: Robert Lerner

Data Quality by the Numbers:

Best Practices for Managing Business Data

DataFlux A SAS COMPANY

Leader in Data Quality and Data Integration

www.dataflux.com
877-846-FLUX

International
+44 (0) 1753 272 020

The development of data quality technology began because of the need to provide accurate name-and-address data for marketing and sales. Over the years, the technology has evolved to support a wide range of name-and-address needs, well beyond those of the sales and marketing departments. After years of advancements in data quality technologies and processes, organizations of all kinds have implemented the technology to support this critical resource.

Name-and-address data, however, is one part of this resource. Most, if not all, organizations have non-customer data, which is frequently referred to as business data; in fact, some organizations have more business data than they have name-and-address data. But, surprisingly, while some organizations will look to data quality technology to solve data quality problems with their customer data, they don't consider data quality technology to support their business data.

Perhaps this isn't surprising, after all. Most data quality solutions have limited capabilities for business data – few solutions, for example, can analyze business data and generate rules that can enforce organizational standards on the data. Furthermore, business data can be extraordinarily complex, which in many cases requires a different approach to the data than is generally used for name-and-address data.

Nevertheless, data quality technology currently exists that can handle business data, and ultimately it is in an organization's best interests to use this technology to support all of their data resources – both customer-oriented and non-customer-oriented. Using data quality technology for names and addresses is only half the battle in terms of protecting one of their most valuable resources – information.

Introduction to Business Data

It is somewhat surprising that many of the organizations that have business data will implement data quality solutions to protect their customer data, but not look to the same technology to protect their other data assets. All data is valuable, but customer data often receives a majority of the concern – and therefore the resources – when data quality projects begin.

Of course, not every organization is like this, but it is true that the use of data quality technology for customer data is far more prevalent than the use of the technology for business data. The reason for this imbalance has as much to do with the organizations that are acquiring data quality solutions as it does with data quality technology itself. Many data quality companies simply cannot support business data to any significant degree. Even some of the companies that profess to support business data can only do so to a limited degree.

This paper will consider the issue of business data and look at some of the most common types of business data, as well as at some of the more esoteric kinds of this data. The purpose is not simply to examine the nearly infinite variety of business data. Instead, this paper will consider the importance and impact of this data and, more significantly, explore the ways that data quality technology can support even the most extreme kinds of business data.

The Problems with Product Data

One of the most common types of business data is product data. Organizations have been having problems with the quality of the product data for as long as there has been product data, but recently there has been a rise in concern about the quality of this data and an organization's ability to manage it effectively. One reason is that, until recently, data quality companies did not do much to support product data. While they were quick to point out the problems inherent in name-and-address data, they said very little about the problems inherent in business data. As a result, many organizations weren't aware that this technology could be useful for this kind of data.

The problems associated with product data were also obscured by the publicity generated by hot applications as customer relationship management (CRM) and customer data warehouses. These applications promised the world, but when the world collapsed because of the poor-quality customer data populating these systems, the media publicized these failures. So, the concerns related to poor-quality customer data got a lot of attention, while the concerns related to poor-quality business data got less media hype. Not surprisingly, data quality companies began hyping their own ability to prevent the failure of CRM and related applications.

But things are changing. Some data quality companies are now touting their ability to handle all kinds of data. At the same time, CRM and other customer-centric solutions have lost some of their luster, and so the media is now paying attention to a broader spectrum of data quality problems. Equally important, the economy, increasing globalization, mergers and acquisitions, regulatory compliance, and other issues that cross geographic boundaries have highlighted the problems associated with poor-quality business data.

The economy, for example, has forced organizations to gain greater control over their inventories, suppliers and reporting, while mergers and acquisitions are bringing diverse, incompatible product data into organizations. This consolidation merely accentuates problems that already exist with their own product data prior to the merger.

Many organizations are also learning that enterprise resource planning (ERP), supply chain management (SCM) and other enterprise applications have not solved these issues. Many organizations mistakenly believed that these applications would eliminate most data quality issues relating to product data. This was akin to the confidence that many organizations placed in CRM. But these applications were never designed to eliminate problems associated with poor quality data and, as a result, many organizations have become disillusioned because their applications can't handle incompatible formats, duplicate data and a range of other data quality problems.

Finally, many organizations today simply have more product data than ever before, which compounds every one of their data quality problems. As data grows exponentially, so do the opportunities for inconsistent, inaccurate and unreliable data. With all these data quality problems, it's easy to see why poor quality product data is topping the agenda of many organizations.

Case Study: Feeling the Effects of Bad Product Data

Manufacturing companies are particularly aware of problematic product data. Many of these companies have reported problems associated with duplicate product numbers, obsolete product numbers, and inconsistent product descriptions or attributes. And these problems exist across the organization, impacting every level of the operation. The result of bad product information can lead to:

- Poor control over inventories
- Inaccurate tracking of product shipments
- Inaccurate planning or forecasting for inventory and raw materials
- Inadequate control of the supply chain
- Ineffective operational analysis

Some time ago, a well-known manufacturer of bedding products had a problem with the accuracy of its finished stock SKUs, and this problem undermined the company's ability to manage its financial health. On the whole, the company's SKUs were fairly accurate, at least in the sense that the SKUs followed acceptable patterns and, when input into computer system, pulled up what looked like reasonable costs for the items in question.

This changed when the company's sales department needed a better way to track the flow of goods being shipped to specific customers. The sales department realized that if each customer had a unique SKU for a particular item, then they could easily track the items shipped to their customers. As a result, the department began to add a letter suffix to the end of each SKU that would relate to a specific customer. For instance, in the fictitious example "XYZ12345678," the sales department would add a "z" to the end of the SKU – "XYZ12345678z" -- to represent an item that went to a specific customer. While this created a unique identifier that helped the sales department track shipments to specific customers and understand what items were selling best in different regions, it also wreaked havoc with the company's financial system.

The company used a standard cost system. Whenever a finished stock SKU was entered into its computer system, the system would automatically apply a "standard" cost to this item. The standard cost is simply a predetermined dollar value (in terms of material, labor, and overheads) that reflects what it should cost the company to make the item. The difference between what it should cost the company and what it actually costs to manufacture the item falls into variance accounts – though in a world without delays, shortages and errors, the standard cost and the actual cost should be the same.

The problem here is that when the system detected a unique character, or an SKU that it didn't recognize (and didn't have a standard cost for), it didn't reject the SKU as one might have expected. Instead, it accepted the SKU and applied a standard cost of $1.00. For example, the company may have sold a mattress, which cost the company about $300.00 to manufacture. But when the suffix was added and the once "legitimate" SKU was now unrecognizable by the system, the system would apply a cost of $1.00 to the mattress – a cost that was far less than anything the company produced.

The difference between the $300.00 it took to manufacture the mattress and the $1.00 that the company's system applied to item fell into the company's variance accounts. Variances, it should be noted, are not a good thing, because they hurt a company's ability to track its costs accurately. And, they suggest that something is wrong with the way that the company is determining and accounting for the cost of producing something. A resolution of variances is critical, and the company was forced to spend critical time and effort reconciling these variances and related problems.

Interestingly, the company stated that once the altered SKU number was in the system, the suffix could not be removed and the altered SKU could not be cancelled. The company's accounting and clerical staff was also forced to spend time and effort going into the system and apply costs for each of the items that had been entered incorrectly.

The problems associated with poor quality product data aren't exclusive to financials. Organizations will typically have problems in determining the cost of raw materials and the best supplier for these materials, or participating effectively in trading networks. Trading networks require accurate, synchronized and up-to-date product information to be effective. If a partner in the network has inconsistent product data, then it becomes difficult for both the organization and the other members of the network to compare prices between partners – and select the best supplier for a given product. Trading networks are becoming increasingly important, especially since giant retailers are requiring their partners to participate in these organizations. Accurate product data, therefore, is crucial to both the partnerships and the trading network itself.

The problems aren't exclusive to manufacturing either. Online retailers, for instance, can have serious problems with customer fulfillment without high quality product data. If the descriptions of the products on the retailer's site are inconsistent or inaccurate – or if the quantities of the products available are inaccurate –incomplete orders, delayed shipments and frustrated customers are likely. In fact, if the quantities are incorrect, the retailer itself may discontinue a product that could still generate business or promote a product that doesn't interest customers.

Beyond Product Data: Numbers

Product data is only the tip of the iceberg in terms of the variety of business data that organizations today must handle. Business data covers everything from product data to operational data to VIN numbers to biological data. In fact, some of the most interesting business data today involves biological data. Biological data is not only extraordinarily complex, but frequently can be understood by only a few people with specialized skills.

The following is a single line from the DNA sequence of a fruit fly:

2041 ttcaggactt caaggatatt ggcaagggag cagcattcta cattacagcc acagtgacaa

This is only one line in thousands of lines in this sequence (line 2041, to be exact), each of which is composed of nothing more than a varying series of four letters (a, g, t, and c, which respectively stand for the nucleotides adenine, thymine, guanine, and cytosine). The complexity of this data presents real challenges to both researchers and those tasked with preserving the integrity of the data. Still, this data requires the same care that any other business data requires, and complexity does not obviate the need to ensure the accuracy of the data both now and over time. Interestingly, problems in the sequence (whether through human error or genetic mutation) could impact scientific study (a human error could obscure a line that may be crucial to study) – and potentially affect human beings, since this data is studied mainly in terms of how such DNA sequences impact human diseases, health, and so forth.

But while this example may seem a bit esoteric in terms of what most businesses would typically expect to grapple with, it nevertheless reflects what might be called more or less typical business data – that is, data that has a specific, identifiable structure.

In many organizations, there is an increasing concern with data that lacks a distinctive structure – data that is, in a sense, just as complex than the most mind-boggling DNA sequence. This kind of data, which is generally thought to be inappropriate for data quality technologies and processes, might be called for lack of a better term numerical data. Numerical data such as trends, statistics, ratios, etc. is just as critical as any other data in an organization. While there are indeed limits in terms of the data quality technologies that can be applied to the data, it is nevertheless crucial to ensure the accuracy, completeness, and quality of this data. Significantly, this data is just as crucial to many organizations as customer names and addresses.

Trend data, for example, is important to many organizations. These organizations have historical data on the buying trends of their most important customers. The trends are important in terms tracking the value of a customer, determining the level of opportunity that the customer represents, following the loyalty of the customer, and so forth. Trend data can also highlight a potential problem in the order/fulfillment process. If, for instance, orders from customer X fall below 10% of historical levels, then the violation of this trend could raise flags that something is amiss either in terms of the order amount or the customer's relationship with the organization.

Other types of numerical data such as statistics or ratios can be just as critical to organizations as trend data, especially in terms of ensuring its quality. Insurance companies, for example, leverage statistics to determine coverage and risk. Statistics such as population density, crime rate, accidents, age, smoking and health are used to develop codes that are then used in conjunction with customers to determine the associated risk of the customer. Each insurance company has an acceptable level of risk for insuring particular customers.

The accuracy of both the statistics and the code are critical, because poor-quality data could expose an insurance company to unacceptable risks. At the same time it could also lead to over/undercharging customers, leading to customer dissatisfaction.

Insurance companies have also determined rates and insurability based on ratios. A company may look at the number of drivers in a particular household and compare this to the number of points or demerits per driver in the household. Essentially, the company will take the number of points in the household, divide the points by the number of drivers in the household, and if the ratio is, for example, greater than .8, the company will flag the insurance records of the members of the household. In this case, the accuracy of both the data determining the ratio and the ratio are significant and the quality of either can impact both the company and its customers.

On the other hand, mortgage companies frequently rely on numerical controls to ensure that the dollar amount of loans going out and the dollar amount of the payments on these loans is correct. In a mortgage company's billing system, the sum of mortgage payments due should theoretically equal the sum of payments to these mortgages. When the two are not equal, there could be a number of issues that require attention – customers paying more or less than the expected amount, customers paying or not paying late fees, etc. – especially when there is inaccurate or incomplete data somewhere in the billing system,.

Despite the variety of this data and the fact that some of it has not traditionally been considered subject to data quality technologies, all data requires the same care and protection to ensure that it remains at the highest quality possible. The problem with some of this data is that even though its quality is essential for any number of reasons, not all data quality technologies are able to support it. But this is not to suggest that there are not data quality technologies that can support the data. These technologies are now available.

Attacking the Problem

Despite the difficulties that business data poses, a number of organizations are taking steps that directly address the data quality issues of business data. Unfortunately, business data poses challenges that are not easy to resolve. Product data can be particularly challenging because it can be highly complex. Some products, for example, can have hundreds and even thousands of attributes. This complexity is multiplied when large organizations have thousands or even millions of products, each with an amazing array of attributes (e.g., dimensions, weight, materials, etc.).

Additionally, product data, unlike customer data, is frequently proprietary and lacks industry standards or standard definitions. A mattress coil, for example, is likely to be unique to a particular manufacturer, despite the fact that it looks superficially like any other mattress coil in the industry. If nothing else, each manufacturer that produces a bedding coil will likely have its own method of producing the coil or will produce the coil for a specific product.

Some organizations have sought to standardize on a particular vendor's applications to create better product data. In theory, this would ensure that all product data formats are consistent. Unfortunately, this tends to be impractical, especially for large organizations that are geographically dispersed and would have to replace a variety of existing applications to ensure this consistency – assuming, of course, that the vendor's various applications are actually compatible with one another. Once again, the applications themselves are not designed to correct data quality problems such as duplicate product records or inconsistent product information.

RFID has also been thought to offer benefits in terms of ensuring consistent and accurate product data. Of course, RFID technology cannot cleanse data, but it can offer advantages if the data starts out accurate, consistent and timely, because it eliminates manual input. However, some of the virtues of RFID – high-speed, high-volume data transfer – could have disastrous effects if the data being transferred is of poor quality, since it would speed the transfer of data errors throughout the organization.

More effective have been industry standards such as commodity coding systems like UNSPSC (United Nations Standards Products and Services Codes) and eCl@ass. These coding systems offer some real advantages, since they provide consistency in terms of product numbering and product descriptions. For all organizations that standardize on UNSPSC, the code 10122101, for instance, has both the same commodity description and meaning to every organization, in this case "Pig Food." There is no confusion in this case over what constitutes pig food, so every organization supporting this code can compare prices between various pig food suppliers more effectively.

Such standards eliminate many product data problems, and they make the correction of product data much easier (similar to applying USPS address standards with customer data). Nevertheless, they don't create high-quality product data. Organizations are in the same position before adopting one of these systems, since they need to cleanse their data and standardize on the coding system to leverage it effectively. Furthermore, they still need to implement technology and processes to ensure that product data continues to meet commodity code standards over time.

There are also a range of standards or methodologies, such as ISO 9000, which can ensure that product data is of the highest possible quality. While these standards are valuable, they also have limitations that can be addressed with a combination of methodology and technology. But it is crucial to recognize that good product data is just as crucial to an organization's health and success as its customer data.

Using Technology to Solve Business Problems

A more effective method of protecting and managing the quality of business data is to leverage data quality technology. However, the question of how to apply data quality technology to business data is not a simple one, as it often is in the case of names and addresses. Essentially, some business data can be protected using all of the “common” data quality technology, while some data requires a subset of this technology.

As a rule, using data quality technology for names and addresses is easier, because this data tends to be less complex than business data. More universal standards exist, and there are even certifications for these standards (USPS has CASS certification, for example). However, business data is difficult to handle because, unlike names and addresses, it generally lacks cultural and governmental standards – standards that cut across organizations and industries and, to some extent, political boundaries.

Obviously, there are standards governing commodity codes (e.g., UNSPSC, eCl@ss), and even DNA sequences following rigid conventions that govern their format. Indeed, if there weren’t such standards for DNA sequences, research on DNA would be practically impossible and whatever research was done would virtually be incommunicable.

But unlike name and address data, business data tends to be proprietary. A manufacturer’s product codes, for example, are usually exclusive to the manufacturer, and therefore they may not be representative of product codes for similar products from other manufacturers. Because of these problems, finding an off-the-shelf solution geared to business data can be quite difficult.

However, some data quality technology can be customized to support a broad variety of business data, and in most of these cases the same capabilities – data profiling, data quality, data integration, data enrichment and data monitoring – that govern customer names and addresses can support virtually any data. Depending on the data, this customization process is very similar to the customization that needs to take place to ensure that a customer name or company name is consistently represented throughout an organization.

Consider the example given above regarding finished stock SKUs. The SKU is formatted as follows: XYZ123456789. The letter prefix designed a specific kind of product (in this example, a mattress or box springs), while the numbers that followed detailed everything from the underlying materials that made up the product to the color of the fabric on top of the product.

If data quality technology can be customized to recognize XYZ as a mattress SKU, ZYX as a box spring SKU, and any other combination of letters as incorrect, then the technology is capable of cleansing and standardizing the prefix. And if the numbers also follow certain restricted patterns – for example, if the first two numbers represent the type of coils inside the unit, the second two numbers represent the type of material surrounding the coils, and so forth – then these patterns can be profiled, cleansed and monitored, with the manufacturer’s internal standards or business rules governing the representation of the data.

Companies can also manage product descriptions contained in the SKUs to ensure that only certain WIP (works in process) SKUs can be rolled up into the finished stock SKUs. In the end, data quality technology could have prevented the situation that occurred when invalid suffixes were added to the finished stock SKUs in the earlier case study – in real time and at the point of entry.

Ultimately, data quality technology can address inconsistencies in the data itself. It can also discover and rectify the internal processes that led to bad or inconsistent data. Even when this data is far more complex than common name and address data – many product descriptions, for instance, can be long and enormously complex – the right data quality technology can profile, parse, cleanse and monitor it.

When Data Defies Technology

What is it about business data that either lacks these standards or defies them? This is especially true with trends, statistics, and numbers. While the underlying data that makes up the trends, statistics, and numbers can be subject to the data quality technology noted above, it is also true that the trends, statistics, and numbers themselves cannot be subjects to the same range of data quality technology. In fact, there is a very limited set of data quality capabilities that can support this data.

Consider mortgage origination fees. Mortgage origination fees are the fees paid to the company that has originated a loan, and the fees typically cover the costs associated with the creation, the processing and the closing of a mortgage.

Typically, mortgage companies will charge around one percent of the total loan amount for origination fees. Now, the amount of this fee is significant to mortgage companies and to borrowers alike, and so ensuring that this percentage is accurate or consistent is critical. Unfortunately, while this data needs the same “protection” as any other kind of data, it really can’t be cleansed or enriched. But while data quality technology can profile this data to discover the rules that govern loan origination fees, it cannot cleanse or augment this data. Instead, it can monitor the data to ensure that these fees don’t fall outside of these acceptable limits.

Monitoring the business rules to ensure that mortgage origination fees don’t exceed this percentage will go a long way to ensure that fees are correct and that the mortgage company is not over- or undercharging its customers, and that its customers are being charged an appropriate amount, according to the terms of the company’s policy. Indeed, violations to the business rule that governs this fee will alert the mortgage company that something is amiss in its data or business processes.

Automobile VIN numbers offer an interesting analogy. VIN numbers are 17 characters long, and each character represents automobile characteristics that can be used to identify a specific automobile. For example, the first number represents the country in which the automobile was manufactured (e.g., numbers 1 and 4 represent the USA, 2 for Canada, 3 for Mexico, etc.), the second number represents the manufacturer (A for Audi, T for Toyota, G for General Motors, etc.), the third character the automobile type, the fourth to eighth characters a variety of body features (style, engine type, model, etc.), and so forth. Because of the variability, it is difficult if not impossible to apply the full range of data quality processes to ensure that these numbers are tracked accurately.

However, this doesn’t mean that the quality of VIN numbers cannot be addressed or subjected to some data quality technology. In this case, data quality technology can monitor the VIN numbers to ensure that they fall within an acceptable range of letters and numbers, and that the placement of the numbers and letters follows whatever rules govern the characters. Monitoring the data, or the rules governing the VIN numbers, will ensure that only numbers are the first character, letters are the second character, and so forth. When a letter is placed in the first character position, for instance, the monitoring technology can then send out an alert that something is amiss with that particular VIN.

Finally, DNA sequences themselves can pose problems beyond simple complexity. While the data has a standard format, this format is ultimately problematic because it does not follow a standard format like an address, which has recognizable elements and which can be parsed for cleansing. Essentially, the data has a standard format in terms of line length and constituent elements (the four repeating letters, or the nucleotides).

For now, some data quality technologies can support these very features, which would eliminate human mistakes and computer errors and ensure that greater part of the value of the data is preserved and protected. There are scientific rules that govern the format of the data, and these rules can theoretically be monitored by the data quality solution to allow scientists and researchers to take proactive steps to protect the data whenever a rule has been violated. While this may seem to be rather minimal protection for the data, it is in fact significant protection, since it is not uncommon to find errors in data like this.

Where to Begin

Most organizations have some sort of business data. This data could be part number or biological data, or numerical data such as trends, statistics or ratios. Given the prevalence of this kind of data, it is important that organizations protect this data just as they might protect their name-and-address data.

Not all data quality technologies and processes work equally well with business data. But there are data quality technologies and processes that can be leveraged to ensure the ongoing quality of business data.

Organizations, therefore, should begin a data quality program with idea that all of their data – not just customer-centric data – should be protected and supported. Organizations should begin a data quality program, one that includes technology as well as processes, data governance and best practices, with the idea that it will support all of their data.

They should begin this program with the understanding that various data quality technologies and processes will be needed for differing data. As discussed in this paper, data cleansing isn't appropriate for statistics, although technology that is capable of monitoring data rules certainly is appropriate.

This program should ultimately constitute an organization-wide effort to protect all of its data. Once this is understood, the organization can then begin to take the appropriate steps to acquire technology and set up internal processes that will ensure the quality of all of its data.

Ultimately, it is critical to understand that all organizational data requires technology and processes to protect and keep the highest quality. It is also important to recognize that simply by supporting customer-centric data, organizations will not get the utmost benefit from this strategic resource. Organizations that only take one step are really missing the point in terms of the importance and benefits of high quality data. Not only are they missing the point, they are also missing some of the benefits of high quality data and, moreover, potentially opening themselves up to other data quality issues. A true data quality effort begins with all organizational data, regardless of distinctions.